Putting the H into NLP

HCI + NLP Workshop

EACL 2021

Marti Hearst

UC Berkeley

April 22, 2021

What does HCI + NLP Mean?

Using NLP to Help People within UIs

Using HCI Techniques to Improve NLP

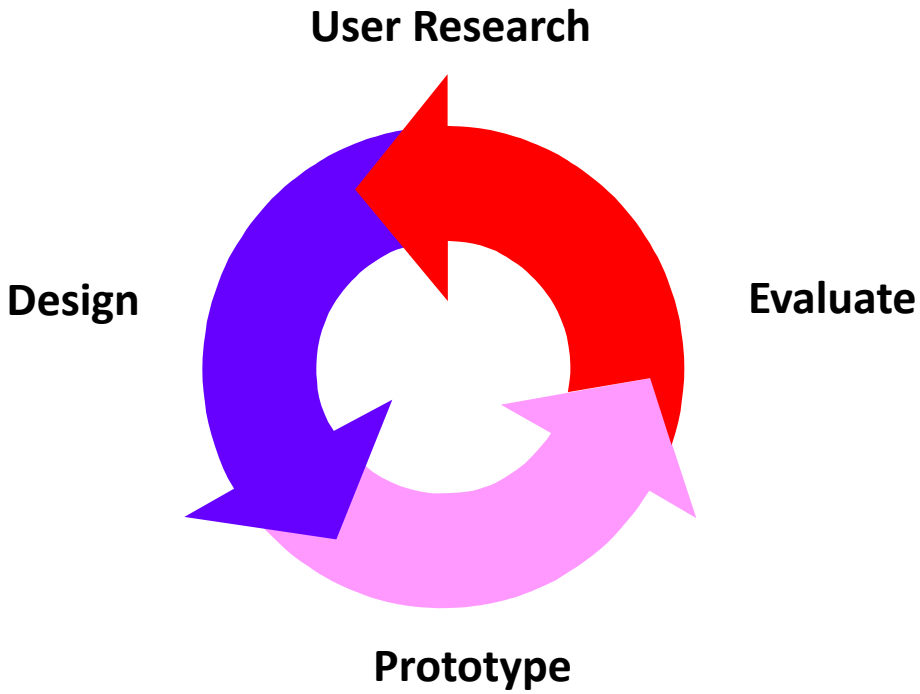
Talk Outline

The User-Centered Design Process

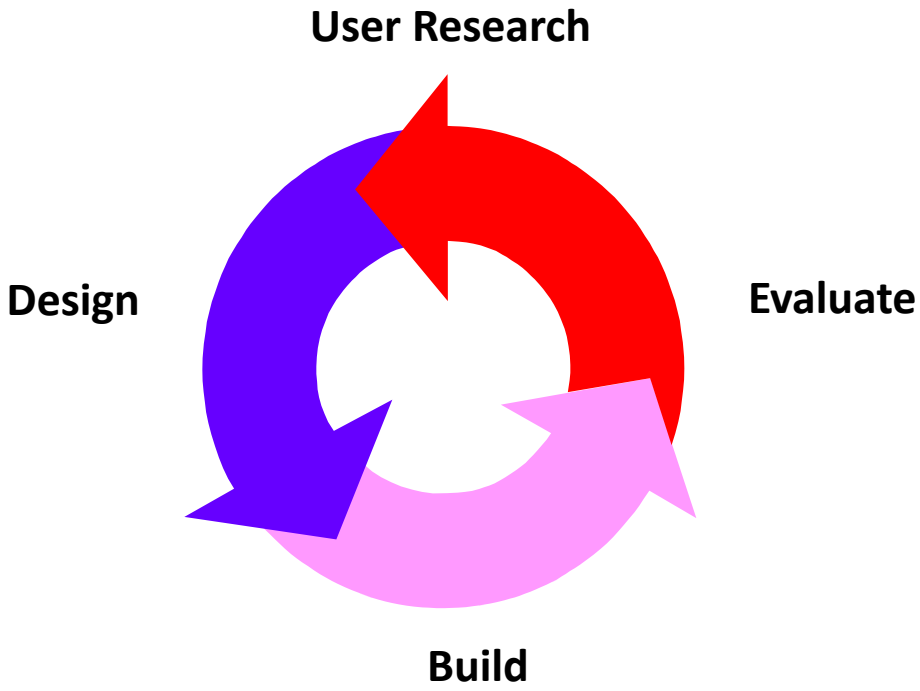
3 Demos

Lessons on HCI + NLP for each

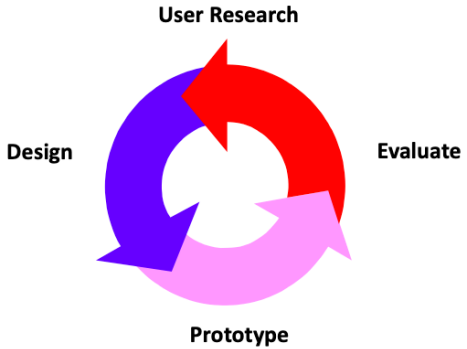
User Interface Design is an Iterative Process



User Interface Design is an Iterative Process



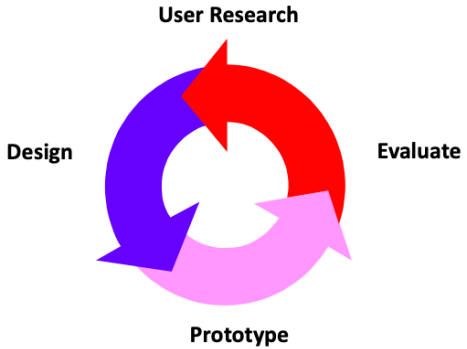
User Research



- Understand
  - Who users are
  - What their goals are
  - What tasks they need to perform.
- Task Analysis
  - Characterize what steps users need to take
  - Create scenarios of actual use
  - Decide which users and tasks to support.

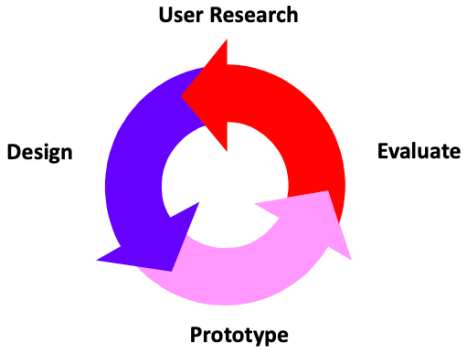
Kim Shephard / Lines of Communication / https://linesof.com/2020/04/12/going-low-tech-with-paper-prototypes/

Why Prototype?



- Experiment with alternative designs
- Fix problems before code is written
- Keep the design centered on the user

EVALUATION



- "Discount" techniques
  - *Pilot tests on prototypes*
  - *Expert evaluations on prototypes*
- Evaluations with participants
  - *Task and interface specific*
  - *Designing these requires care, thought, and iteration*

Typical HCI Process vs Typical NLP Process

HCI:
Identify user need
Develop method to address it
Evaluate method on user needs

NLP:
Identify NLP problem
Develop algorithm
Evaluate algorithm on accuracy, speed

Demo 1:

ScholarΦ

ScholarΦ

Augmented Reader for Scientific Papers

CHI 2021 (to appear)

Marti A. Hearst

Daniel S. Weld

Dongyeop Kang

Andrew Head

Kyle Lo

Raymond Fok

Berkeley UNIVERSITY OF CALIFORNIA

AI2

UNIVERSITY of WASHINGTON

ALFRED P. SLOAN FOUNDATION

Motivation:

Have you ever struggled to keep track of notation or acronyms when reading a paper?

$$\frac{1}{T}\sum_{t=1}^{T}\Big[\sum_{f=1}^{F}\log P(y_{ft}^{role} \mid \mathcal{P}_G, \mathcal{V}_G, \mathcal{X}) + \log P(y_t^{prp} \mid \mathcal{X}) + \lambda_1 \log P(\text{head}(t) \mid \mathcal{X}) + \lambda_2 \log P(y_t^{dep} \mid \mathcal{P}_G, \mathcal{X})\Big] \quad (7)$$

Strubell et al., EMNLP 2018

What does y stand for again?

Motivation:

Have you ever struggled to keep track of notation or acronyms when reading a paper?

He et al. (2018)	84.9	**85.7**	85.3	84.8	**87.2**	86.0	73.9	**78.4**	76.1
SA	85.78	84.74	85.26	86.21	85.98	86.09	77.1	75.61	76.35
LISA	**86.07**	84.64	**85.35**	86.69	86.42	86.55	78.95	77.17	78.05
+D&M	85.83	84.51	85.17	**87.13**	86.67	**86.90**	**79.02**	77.49	**78.25**
+*Gold*	*88.51*	*86.77*	*87.63*	—	—	—	—	—	—

Table 1: Precision, recall and F1 on the CoNLL-2005 development and test sets.

Strubell et al., EMNLP 2018

What does D&M stand for again?

ScholarPhi Demo Video

Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols

Andrew Head • Kyle Lo • Dongyeop Kang • Raymond Fok
Sam Skjonsberg • Daniel S. Weld • Marti A. Hearst

Berkeley UNIVERSITY OF CALIFORNIA
AI2
UNIVERSITY of WASHINGTON
ALFRED P. SLOAN FOUNDATION

HCI + NLP Tip Appears Here

Elaboration appears here

Usable Interfaces Require Multiple Rounds of Pilot Testing and Refinement

Four rounds of piloting

(In this case, the core idea was always positively responded to.)

Refined for months, continual feedback from relevant users

Studied the relevant HCI literature (reading, hypertext, etc.)

Real Applications Shed Light on Tough, Underexplored NLP Problems

Definition recognition is far from solved, and under-explored.

Complex, unsolved coordination problems.

Small Details Matter

Logical organization of information

Visible cues for current and next choices

Interactions that sustain flow

Legible text without artifacts

Standards followed

When HCI Techniques Not Used…

"In a live system, presentational details become **disproportionately** important. In our initial deployment, rendered text contained artifacts of the underlying tokenization …

These were no doubt relatively **trivial matters of software engineering**, but in initial informal evaluations, users kept mentioning these imperfections over and over again … distracting them from considering the underlying quality of the [system]."

https://www.groundai.com/project/rapidly-deploying-a-neural-search-engine-for-the-covid-19-open-research-dataset-preliminary-thoughts-and-lessons-learned/1

Demo 2: WordZones

WordZones: Improving Word Cloud Designs

IEEE TVCG 2019

Marti Hearst, Emily Pedersen, Lekha Patil, Elsie Lee, Paul Laskwoski, Steven Franconeri





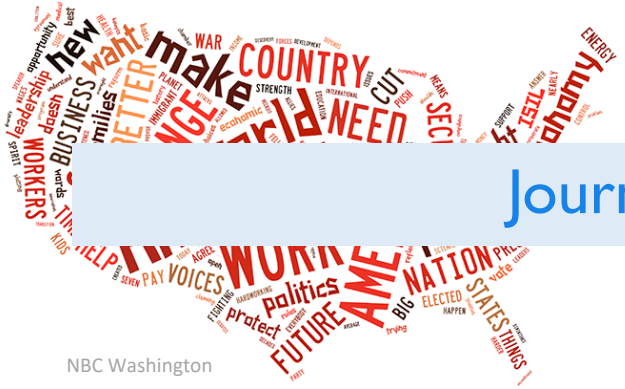

http://www.contrib.andrew.cmu.edu/~dyafei/NLP.html

Word Clouds Are Engaging

They were original intended for fun

http://www.contrib.andrew.cmu.edu/~dyafei/NLP.html

However, today word clouds are frequently (mis)used in …



NBC Washington

Journalism

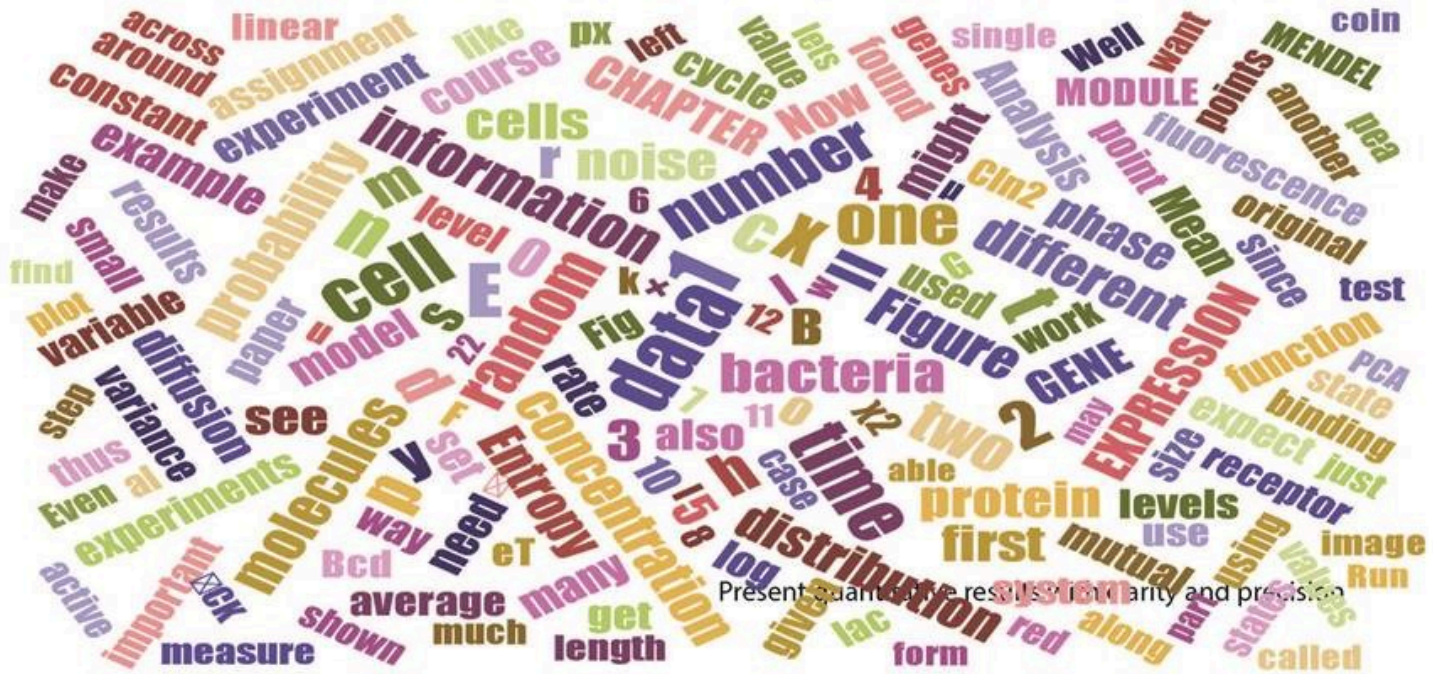
USA Today

Vis 2018

Scientific communication

Vis 2018

What Is This About?



Is it a good way to show a summary of an upcoming course?

But they do not convey the intended message.

"Obama … set out an agenda … turning back the effects of **climate change**, launching a "moonshot" to **cure cancer**, and a grassroots movement to demand **changes in the political system**."

USA Today, January 12, 2016

Hypothesis

Standard word clouds are detrimental to understanding.

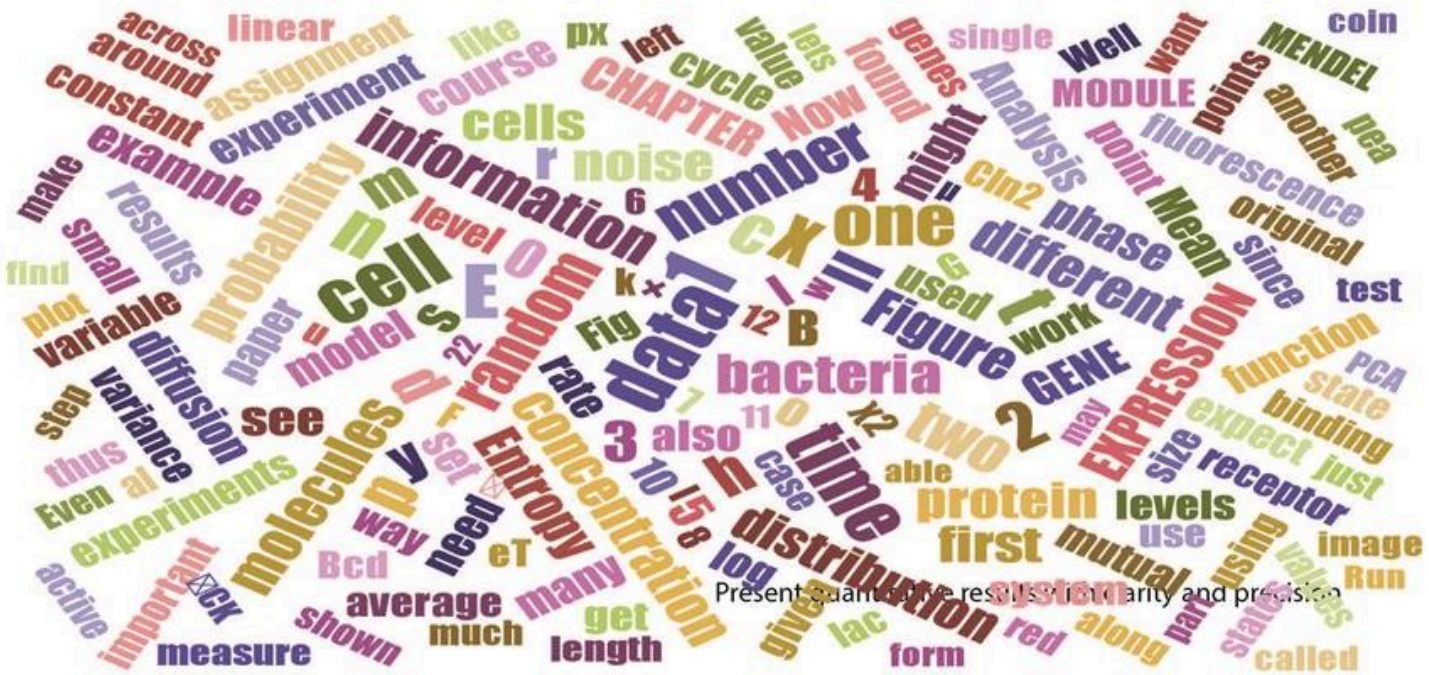
A list of words is better for summarizing a text.

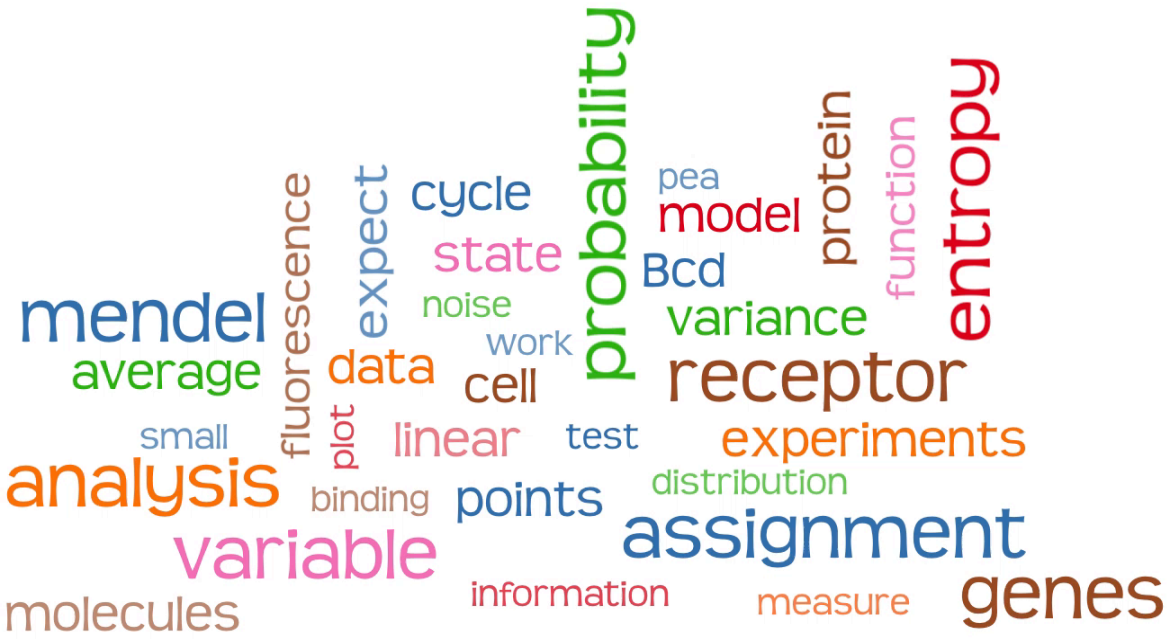
Hypothesis

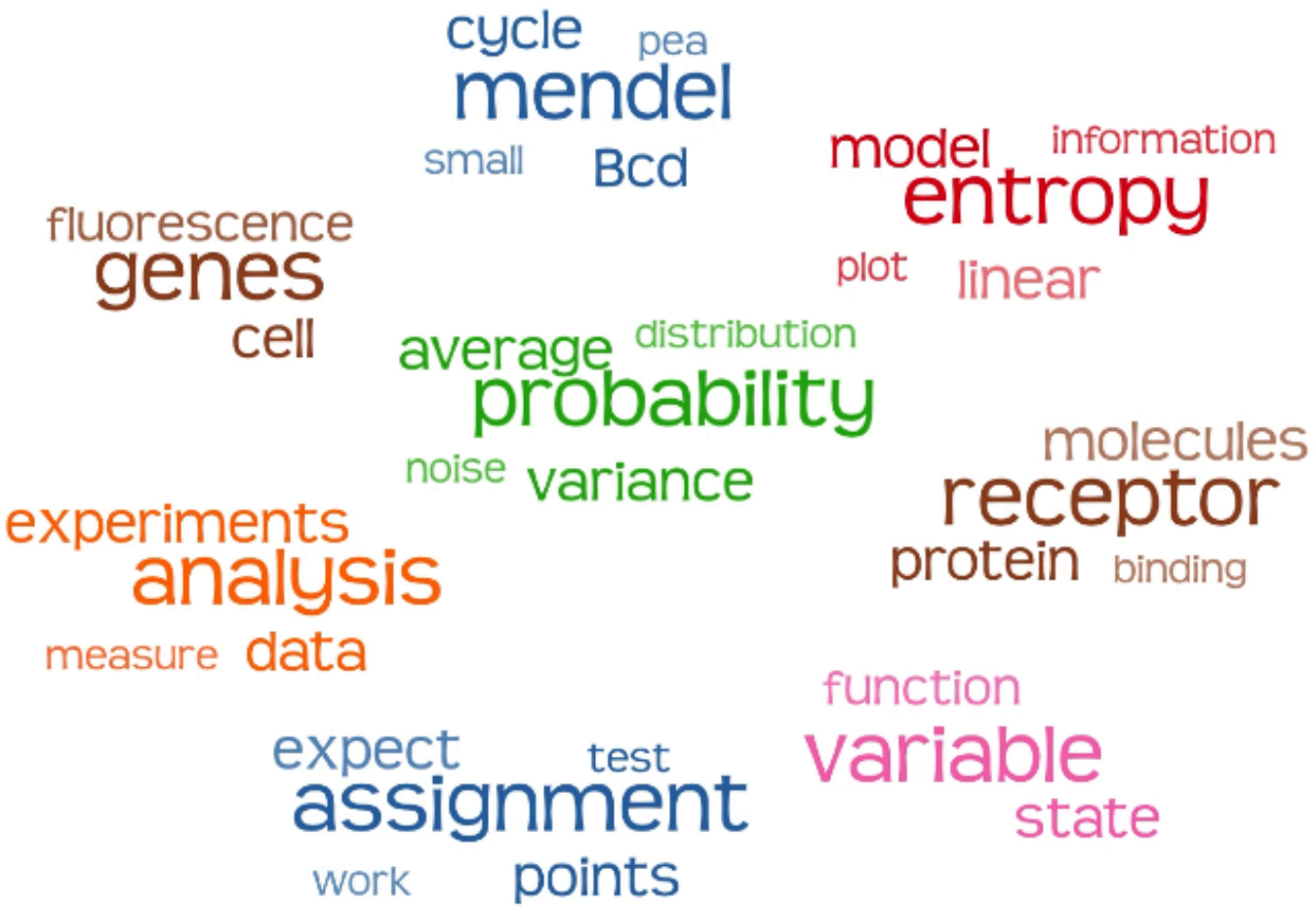
Organizing the words both semantically and visually will **improve understanding** while **retaining engagement.**

How to prove this?

Can we fix this “informational” word cloud?




mendel
average
analysis
variable
molecules
small
fluorescence
plot
expect
data
binding
cycle
state
noise
work
cell
linear
points
information
probability
pea
model
Bcd
variance
receptor
test
distribution
assignment
measure
protein
function
entropy
experiments
genes


cycle
pea
mendel
small
Bcd
model
information
entropy
fluorescence
genes
plot
linear
cell
average
distribution
probability
molecules
noise
variance
receptor
experiments
protein
binding
analysis
measure
data
function
variable
expect
test
assignment
state
work
points

How to Test the Hypotheses? Problems with Prior Evaluations

No shared datasets; no reproducibility

Often no human evaluation at all

Often unfair baselines, and only one baseline

Tasks often do not match underlying goals

"Find the largest word"

A New Evaluation Method

Goal: Determine how well a layout "summarizes" the main topics of a document while retaining engagement

Goal: Reproducible evaluation

Goal: Evaluation the reflects the task

IDEA: "TABOO" WORDS

New Approach: based on the game of Taboo

- Build sets of words that **unambiguously** indicate a category
- The categories simulate the topics of a document
- See how long it takes someone to guess the underlying categories when the words are arranged in a cloud.

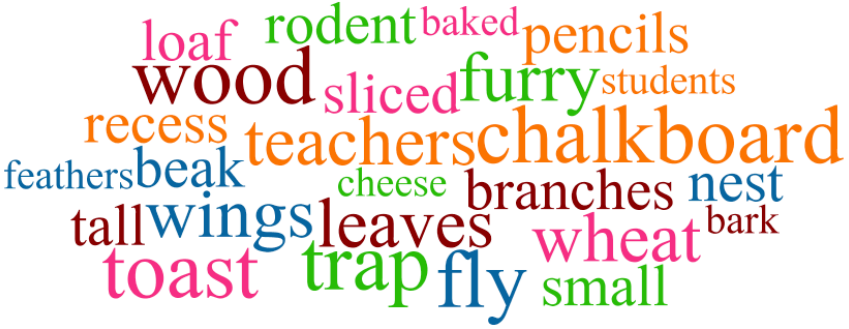
menu waiter dishes tablecloth bill

?

restaurant

95% agreement or higher in isolation with native English speakers
These categories were time consuming to build; required rounds of interation


moo
udder
professors
farm
dorms
students
dairy
milk
majors
campus



How many categories can you name in the time limit?
Can score from 0 to 5.

Column
Single Font
Mono Color

Column
Multi Font
Mono Color

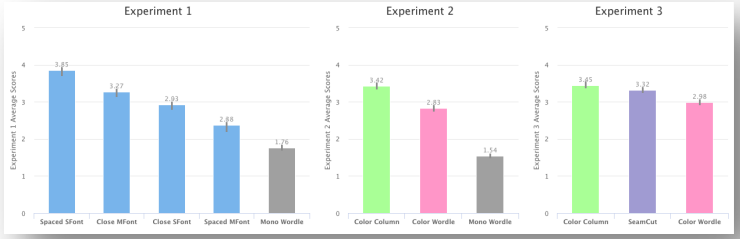
Wordle
Single Font
Mono Color

Wordle
Multi Font
Mono Color

Hold some aspects constant, compare others

Benefits of this Evaluation Approach

- Reproducible
- Allows for fair comparisons across many baselines
- Yields statistically significant results
- Measures some of the underlying goals of the design
  - *(Pair it with subjective evaluations)*



Evaluate with Realistic, Relevant Tasks

Make sure the evaluation task matches realistic use cases.

Observe both subjective responses and quantitative results.

Do Compare Against a Strong Baseline

Compare against a full-featured baseline

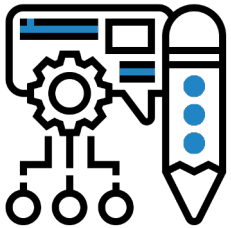
Be sure the available contents are the same

Ideally vary as little between the two designs as possible

Do Reproducible Studies

(This is a lesson for HCI from NLP)

Demo 3:
News Chatbot



NewsLens
A Question Driven News Chatbot

Philippe Laban - phillab@berkeley.edu
John Canny and Marti Hearst
ACL 2020 System Track

Berkeley
UNIVERSITY OF CALIFORNIA

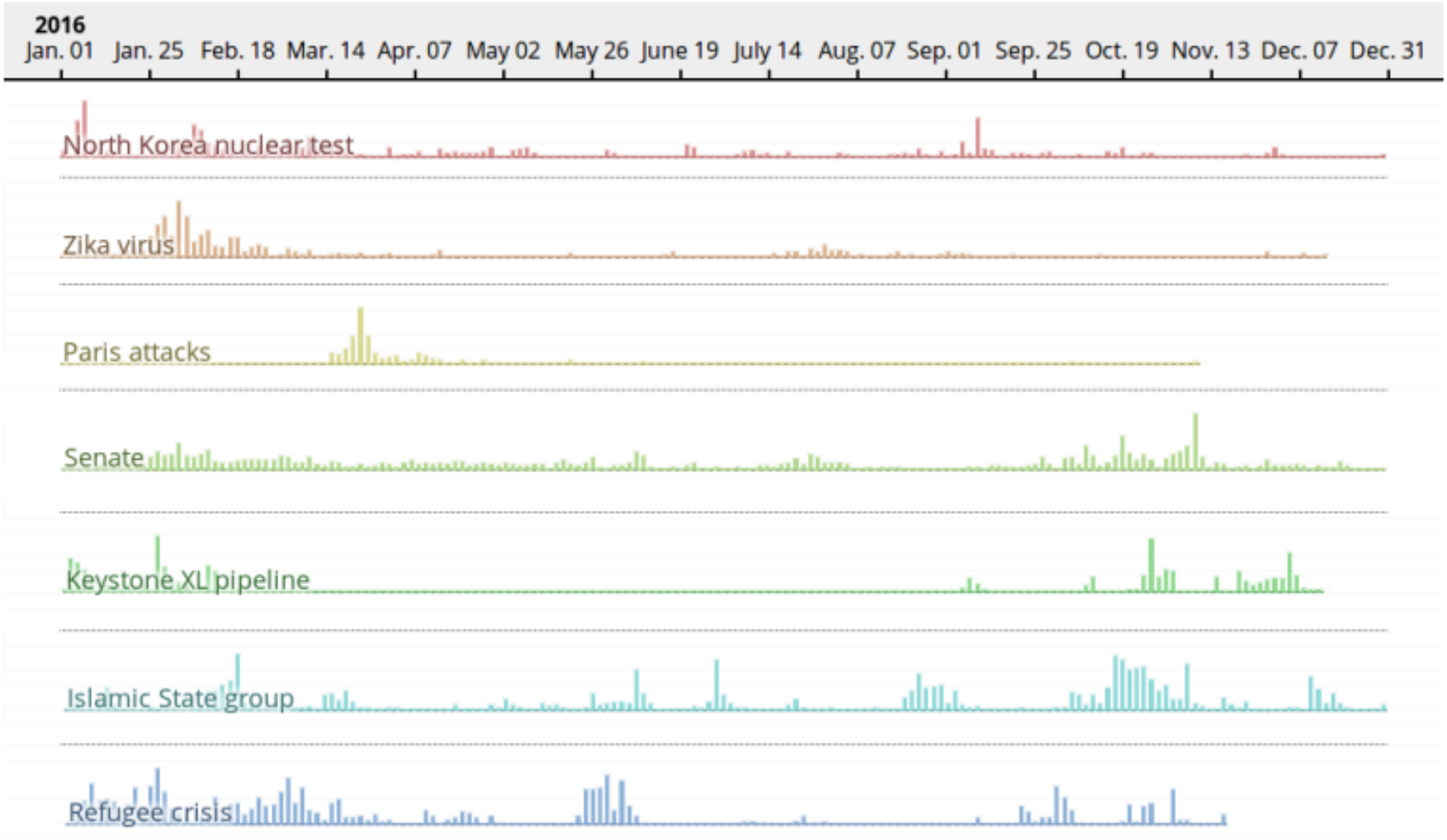
amazon web services™

Bloomberg

NVIDIA®

NewsLens Project

- **Goals**:
  - *Get people to engage in news*
  - *Help journalists*
- **Approach**:
  - *Analyze millions of news articles*
  - *Automatically detect stories, timelines, entities*
  - *Visualize these*
- **Results**: Great NLP research in summarization, simplification, entity recognition, headline grouping. Creation of a highly valuable dataset for NLP
- **Problem**: Cool UI did not engage lay people or help journalists


2016
Jan. 01
Jan. 25
Feb. 18
Mar. 14
Apr. 07
May 02
May 26
June 19
July 14
Aug. 07
Sep. 01
Sep. 25
Oct. 19
Nov. 13
Dec. 07
Dec. 31
North Korea nuclear test
Zika virus
Paris attacks
Senate
Keystone XL pipeline
Islamic State group
Refugee crisis

Cambridge Analytica Scandal

Actors

Search for an actor...

37 director | 36 president | 29 senator | 24 professor

Mark Zuckerberg • 2138 mentions
ceo | executive | chief executive | founder

Aleksandr Kogan • 773 mentions
researcher | professor | scientist | cambridge

Donald Trump • 741 mentions
candidate | president | presidential candidate | us president

Alexander Nix • 550 mentions
ceo | executive | chief executive | chief officer

Christopher Wylie • 494 mentions
employee | former employee | whistleblower | former cambridge analytica employee

Shahmir Sanni • 242 mentions
volunteer | whistleblower | blower | - blower

Damian Collins • 235 mentions
chair | chairman | mp | head

< 1/126 >

Apr. 2018

Apr. 10 2018 (3) Facebook CEO Mark Zuckerberg is about to testify to Congress
Open details

Apr. 06 2018 (3) EU says Facebook confirmed data of 2.7 million Europeans 'improperly shared'
Open details

Apr. 04 2018 (7) Zuckerberg will testify before Congress on April 11
Open details

Mar. 27 2018 (4) Cambridge Analytica whistleblower: Vote Leave 'cheating' may have swayed Brexit referendum - video
Open details

Mar. 27 2018 (5) Facebook CEO Mark Zuckerberg will reportedly testify before Congress
Open details

Mar. 27 2018 (4) Zuckerberg will not appear before UK parliament committee
Open details

Mar. 26 2018 (7) FTC Confirms Facebook Probe Over Privacy Practices
Open details

newsLens Search...

Saudi crown prince X

Actors


Mohammed bin Salman
1490 101 Q


King Abdullah
1083 3 Q227


Mohammed bin Nayef
618 11 Q240


Barack Obama
317 57 Q58


Donald Trump
165 6 Q39


King Salman
141 3 Q44


Ibn Saud
99 0 Q30


Crown Prince
91 0 Q30


Prince Muqrin
86 0 Q55


Prince Sultan
84 0 Q18


Vladimir Putin
71 9 Q16


Prince Mohammed
69 2 Q15

1 / 98

Media

Nov. 07

Nov. 01

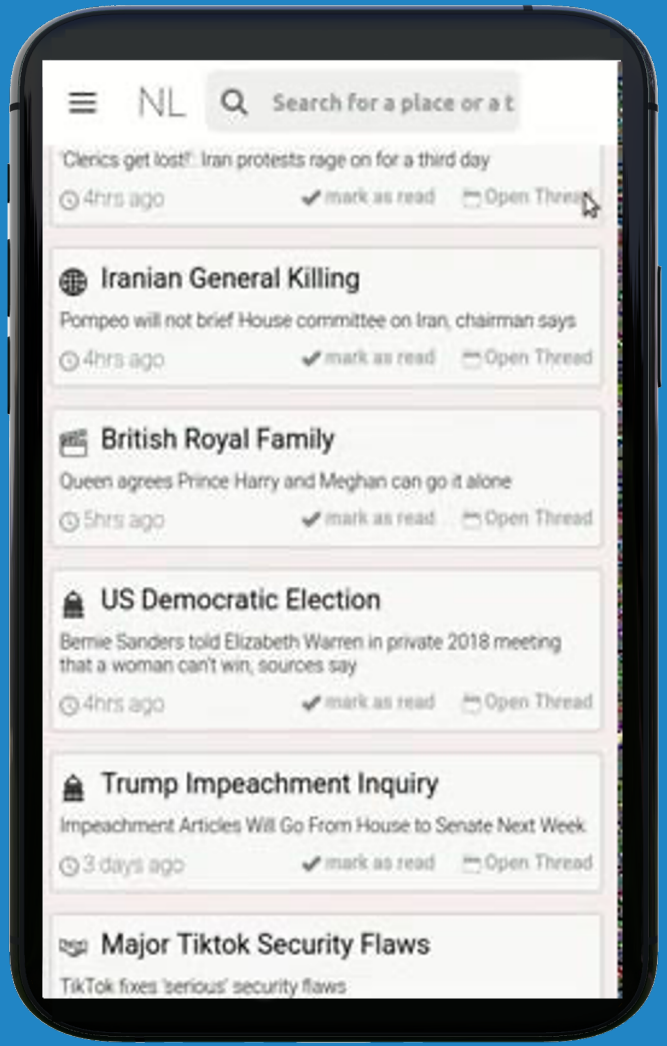
Saudi crown prince

- Saudi Arabia corruption arrests are a bold but risky attempt b...
- Royal purge sends shockwaves through Saudi Arabia's elites
- Senior Saudi royal ousted, princes reportedly arrested in pow...
- Saudi Arabia Arrests 11 Princes, Including Billionaire Alwaleed ...
- Saudi Arabia to begin issuing tourist visas
- Saudi Arabia to allow women spectators in sports stadiums - ...
- Saudi Arabia to allow women into sports stadiums starting in ...
- Saudi Arabia Is Open for Business, but Not Everybody's Buyin...
- The House of Saud is still in denial
- Saudi Arabia Just Announced Plans to Build a Mega City That ...
- Saudi Prince Tells Investors He's Taking on Religious Extremists

New Approach: Newslens Chatbot

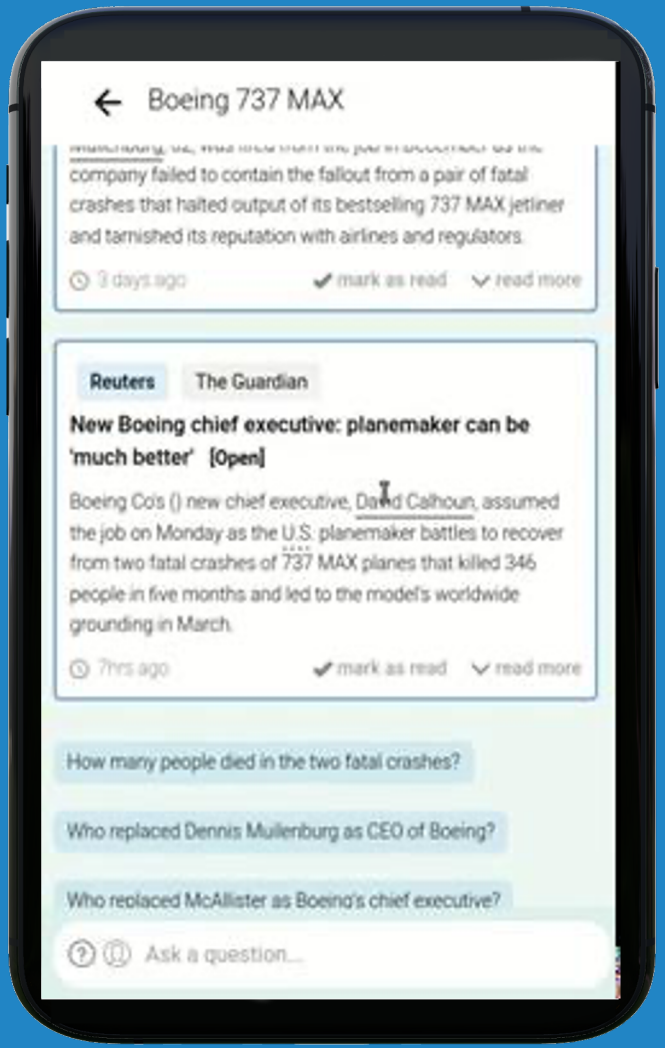
- **Goal**: Get casual news readers to engage more in-depth
- **Hypothesis**: Question-asking will increase engagement

NewsLens Chatbot Demo



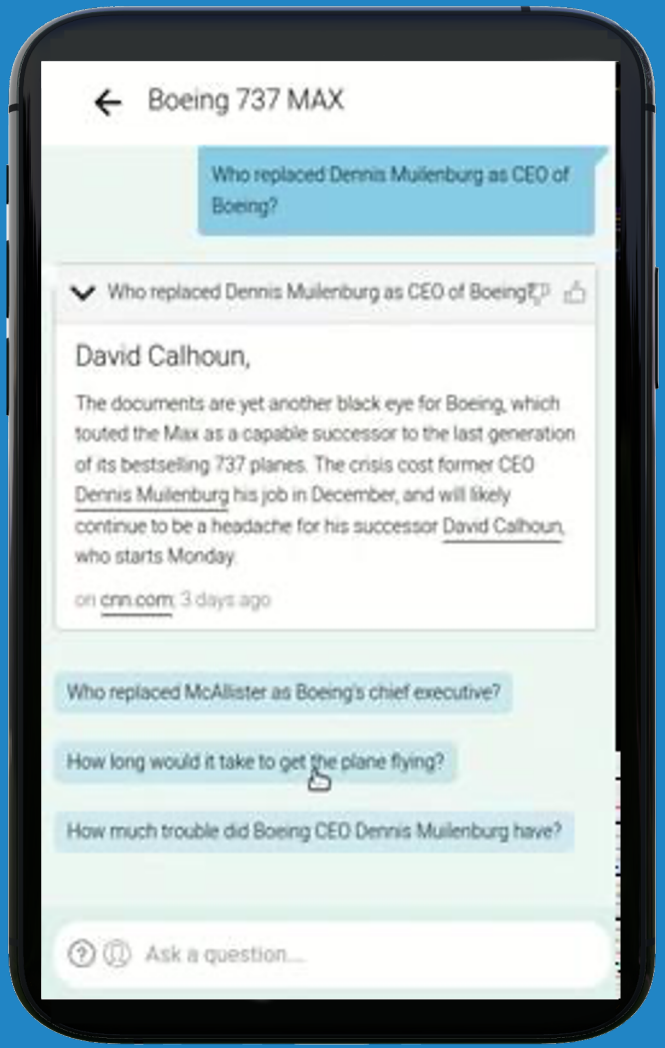
News is organized into *stories*, each forming a *chatroom*.

Example stories:
Iran Plane Crash,
Boeing 737 MAX…



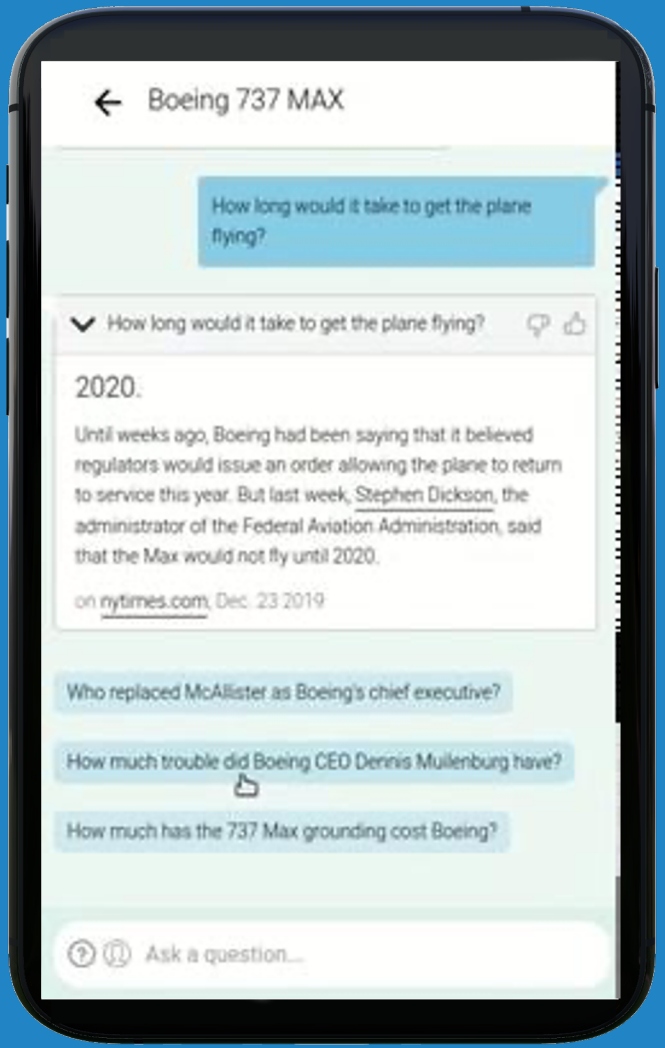
A chatroom starts with a timeline of events.

Each event is composed of multiple sources.

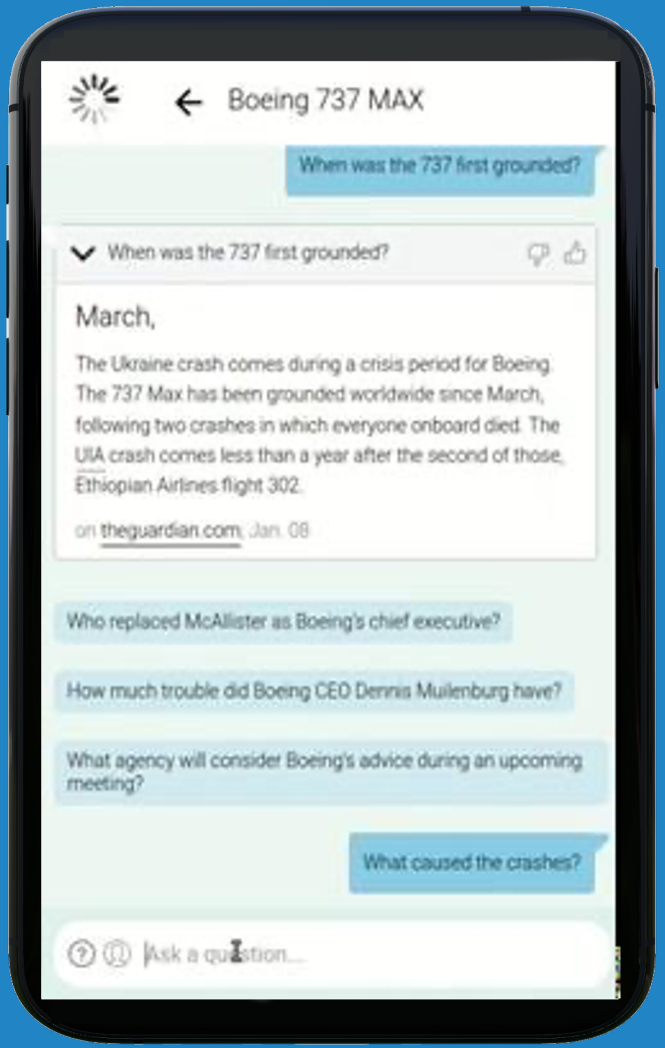


Questions are recommended to the reader.

Questions are updated as information is revealed.



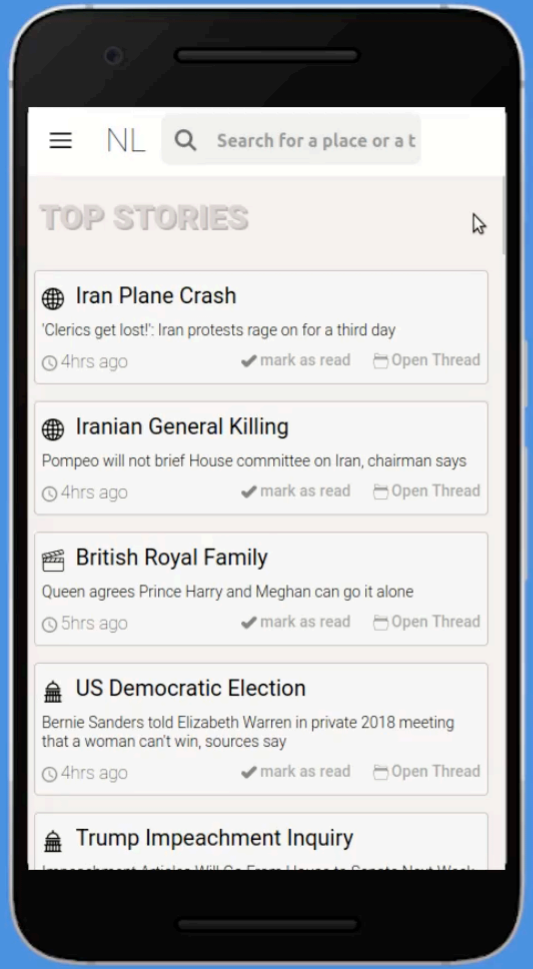
Information is gathered from news and Wikipedia sources.



The reader can ask their own questions.

An extractive Q&A system finds a likely answer.

NewsLens


NL
Search for a place or a t
TOP STORIES
Iran Plane Crash
'Clerics get lost!': Iran protests rage on for a third day
4hrs ago
mark as read
Open Thread
Iranian General Killing
Pompeo will not brief House committee on Iran, chairman says
4hrs ago
mark as read
Open Thread
British Royal Family
Queen agrees Prince Harry and Meghan can go it alone
5hrs ago
mark as read
Open Thread
US Democratic Election
Bernie Sanders told Elizabeth Warren in private 2018 meeting that a woman can't win, sources say
4hrs ago
mark as read
Open Thread
Trump Impeachment Inquiry

NEWSLENS CHATBOT

- **Goal**: Get casual news readers to engage more in-depth
- **Hypothesis**: Asking questions will increase engagement
- **Study Results**: When the chatbot recommends questions, news readers tend to have longer conversations
- **More Recent Work**: An audio podcast version finds that participants prefer a news podcast with Q&A vs one with stories read straight through

Real Applications Can Yield Highly Valuable Datasets for NLP Research

The unique NewsLens collection has enabled cutting-edge research in unsupervised summarization, simplification.

Beware of Cool vs. Usable

Develop something that looks cool.

However, it might not work as an interface for the intended users.

Putting Cool before Usable

"After spending a lot of effort on the [cool NLP problem], the feedback we received from [our users] was quite sobering. Apparently, [our users] are a target audience with needs and preferences quite different from what computational linguists would prescribe."

"To summarize, [the users] opted for a more intuitive presentation style…"

Summary: HCI + NLP

Using NLP to Help People within UIs

Using HCI Techniques to Improve NLP

How To Come Up With a Successful Novel User Interface Design?

- Be sure you've identified a real need
- Put user needs ahead of technology coolness
- Pilot test, pilot test, pilot test
- Small details matter as much as large ideas

How To Avoid Evaluation Errors?

- Don't assess what makes one's own technology look good versus assessing what people need or understand
- Don't measure what is easy to measure versus what matters

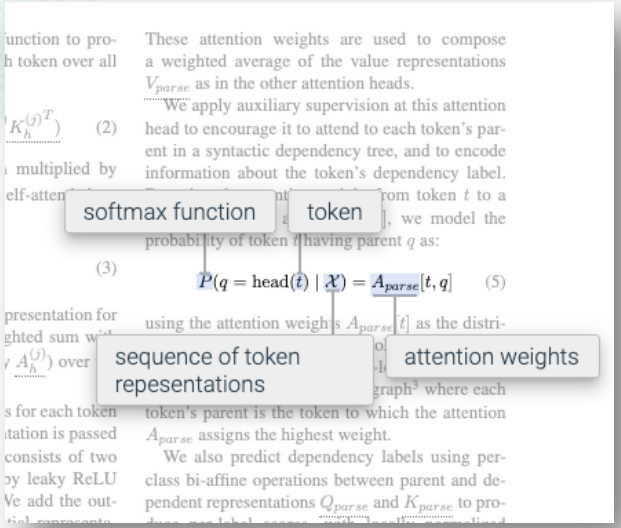
How To Devise a Good Evaluation?

- Spend a lot of time thinking about it
  - *Read other experimental paper*
  - *Think deeply about the underlying goals of the application*
- Pilot test the measurement
  - *Refine it until it is getting consistent results*

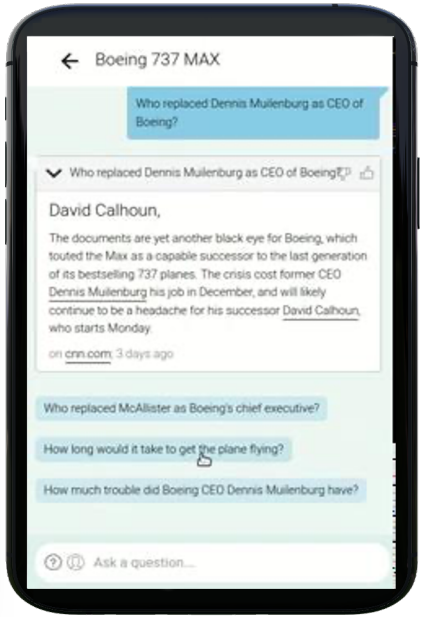
Thank you!

Marti Hearst

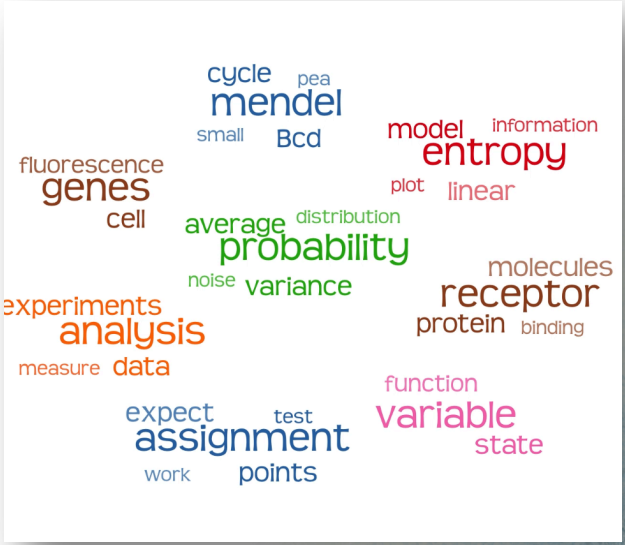
Putting the H in NLP



ScholarPhi



NewsLens



WordZones